ABSTRACT OF THE DISCLOSURE

A system and a method are described for rapidly determining document similarity among a set of documents, such as a set of documents obtained from an information retrieval (IR) system. A ranked list of the most important terms in each document is obtained using a phrase recognizer system. The list is stored in a database and is used to compute document similarity with a simple database query. If the number of terms found to not be contained in both documents is less than some predetermined threshold compared to the total number of terms in the document, these documents are determined to be very similar. It is shown that these techniques may be employed to accurately recognize that documents, that have been revised to contain parts of other documents, are still closely related to the original document. These teachings further provide for the computation of a document signature that can then be used to make a rapid comparison between documents that are likely to be identical.